Customer Churn Analysis Using R

A Major Project

Submitted in partial fulfilment of the requirement for the award of the degree of

Bachelor of Technology

In

COMPUTER SCIENCE AND ENGINEERING

By

KARTIK SETHI (10316210054)
And
MUDEET JAIN (10316210072)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

FACULTY OF ENGINEERING

SRM UNIVERSITY DELHI-NCR

Plot No.39, Rajiv Gandhi Education City, P.S.Rai, Sonepat, Haryana – 131029

MAY 2020

Customer Churn Analysis

Using R

A Major Project

Submitted in partial fulfilment of the requirement for the award of the degree of

Bachelor of Technology

In

COMPUTER SCIENCE AND ENGINEERING

By

KARTIK SETHI (10316210054)
And
MUDEET JAIN (10316210072)

Under Supervision of
DR. PUNEET GOSWAMI

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

FACULTY OF ENGINEERING

SRM UNIVERSITY DELHI-NCR

Plot No.39, Rajiv Gandhi Education City, P.S.Rai, Sonepat, Haryana – 131029

MAY 2020

Department of Computer Science and Engineering SRM UNIVERSITY DELHI-NCR
Plot No.39, Rajiv Gandhi Education City,P.S. Rai, Sonepat, Haryana-131029
http://www.srmuniversity.ac.in

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the project entitled "**CUSTOMER CHURN ANALYSIS USING R**" in partial fulfilment of the requirement for the award of the degree of Bachelor of Technology in Computer Science & Engineering and submitted in the Department of Computer Science & Engineering of SRM University, Delhi-NCR, Sonepat, Haryana, (India) is an authentic record of my own work carried out under the supervision of **Dr. PUNEET GOSWAMI**. The matter presented in this project has not been submitted for the award of any other degree of this or any other Institute / University.

(Signature of the candidate)
KARTIK SETHI
(10316210054)

(Signature of the candidate)
MUDEET JAIN
(10316210072)

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

DR. PUNEET GOSWAMI
Professor

The B. Tech. project viva-voce examination of **Kartik Sethi** and **Mudeet Jain** has been held on _____

Internal Examiner **External Examiner**

Department of Computer Science and Engineering
SRM UNIVERSITY DELHI-NCR
Plot No.39, Rajiv Gandhi Education City,P.S. Rai, Sonepat, Haryana-131029
http://www.srmuniversity.ac.in

CERTIFICATE

This is to certify that the project titled "CUSTOMER CHURN ANALYSIS USING R" is the bona-fide work carried out by Kartik Sethi and Mudeet Jain, students of B.Tech(CSE) of SRM University Delhi-NCR, Sonipat, Haryana-131029 during the academic year 2019-20, in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology(Computer Science and Engineering) and that the project has not formed the basis for the award previously of any other degree, diploma, fellowship or any other similar title.

Dr. PUNEET GOSWAMI
Professor

Dr. Ajay Sharma
HOD CSE (UG)

ABSTRACT

In today's global world, the current technology whatever we are following, everyone is utilizing online applications of various business organizations. So, every organization need a statistical report for analyzing about their existing customers as well as new customers. Based on that they are providing more feasibility and facilities so that they can hold their customers for a longer period. In this proposed paper my topic is Churn Analysis. So, we are using various churn predictive models to provide the statistical report to the organizations.

In telecommunication companies, 'churn' means customer's decision to move from one service provider to another. The competition environment in telecom companies makes their aim is to maintain their customers who are likelihood to leave and earns their satisfaction, so to avoid the problem of churn, they need churn predictive models.

Data mining techniques can be used to build churn prediction model for telecommunication companies to identify churner and non-churner customers because it can extract the predictive information from large databases.

Retaining one customer costs an organization from 5 to 10 times less than gaining a new one. Predictive models can provide correct identification of possible churners in the near future in order to provide a retention solution. This prediction model is based on Data Mining (DM) techniques using R. The proposed model is composed of six steps which are; identify problem domain, data selection, investigate data set, classification, clustering and knowledge usage. The data mining techniques used in this model are Decision Tree, Logistic Regression Model thus working on the software named RStudio.

ACKNOWLEDGEMENT

The Success and Outcome of this project required a lot of guidance and assistance from many people and we are extremely privileged to have this all along the completion of my project. All that we have done is only due to the supervision and assistance and we would not forget to thank them.

We owe our deepest gratitude to our Project Guide Dr. Puneet Goswami, who took keen interest in our project work and guided us all along, till the completion of our project work by providing all the necessary information for developing a good system.

Lastly, we would like to express our deep appreciation towards our classmates our indebtedness to our parents for providing moral support and encouragement.

Kartik Sethi

(Reg. No. 10316210054)

Mudeet Jain

(Reg. No. 10316210072)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 PROBLEM DEFINITION AND SCOPE

Churn is a term used in many companies which is mean loss of customers of the company for many reasons. One of them is the dissatisfaction of customers.

In telecommunication companies “churn” term refers to customer's decision to leave the current service provider and move to other service provider, it can easily happen especially for prepaid customers because they have not any contract same as to post-paid customers.

Churn occurs easily because of the strong and breeding competition environment in services which are providing especially in telecommunication sector. Also, churn can happen for another reason for examples customer's dissatisfaction with services and high cost of these services which can be in another service provider with best quality and lower cost. So, churn became a concerned issue in telecom sector because retaining of existing customer is less costly than acquiring new one.

To identify churner and non-churner customers and understand the reasons of this churn to reduce it, these companies can build churn prediction model for extracting a predictive information from large databases.

The customer churn analysis feature helps us identify and focus on higher value customers, determine what actions typically precede a lost customer or sale, and better understand what factors influence customer spending.

Numerous telecom companies are present all over the world. Telecommunication market is facing a severe loss of revenue due to increasing competition among them and loss of potential customers. Many companies are finding the reasons of losing customers by measuring customer loyalty to regain the lost customers. To keep up in the competition and to acquire as many customers, most operators invest a huge amount of revenue to expand their business in the beginning. Therefore, it has become important for the operators to earn back the amount they invested along with at least the minimum profit within a very short period of time.

1.2 ABOUT THIS PROJECT

In this project, we are going to implement a system in R which analyses Telecommunication data where cluster of nodes will be formed. Here, data is in the form of records of the customers which include telecom training dataset and test data set. By analyzing this data, our system will give output in the form of Graphs, Boxplots etc. about:

- Who are the customers likely to churn?
- What are the factors mainly responsible for churn?
- What type of recommendations can be provided to the organization? Who are the most valuable customers?
- How valid is the churn prediction model?

1.3 OBJECTIVE OF THIS PROJECT

The objective of this problem is to build a churn prediction model which can identify churner customers and non-churner customers. The following project aims to take data from a telecommunication company and use it to train classification models that will be able to predict future churn behavior of customers.

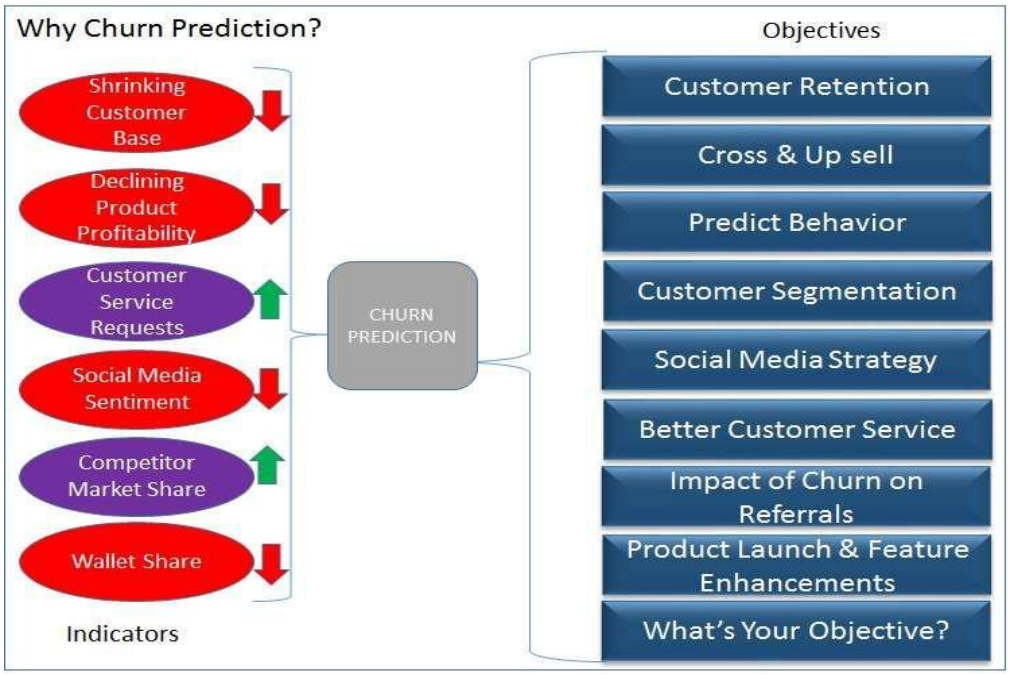


Fig:1 Need For Customer Churn Prediction

1.4 R PROGRAMMING LANGUAGE

R programming language is best for statistical, data analysis and machine learning. By using this language, we can create objects, functions, and packages. We can use it anywhere. It's platform- independent, so we can apply it to all operating system. By using R, we can create any form of statistics and data manipulation. The R language has represented the large dataset in various forms of graphs which provides the outcome in unique pattern visualizations.

R is a powerful statistical programming language which can deal with large dataset and represent it graphically with different parameters and it uses different packages available.

R has some statistical features also:

Basic Statistics – Mean, variance, median.

Static graphics– Basic plots, graphic maps.

Probability distributions – Beta, Binomial.

1.4.1 Business Adoption of R

Many data analysts and scientists use R programming language as a tool to analyse large dataset. Large businesses need open source tools and technologies to analyse their large dataset. Since R is an effective and open source tool, R has slowly became largely used language by various big companies for their large data analysis. Companies such as Google, Genpact, Facebook and various others are incorporating R language in their business processes.

1.4.2 Advantages of R Language

Many quantitative analysts use R as their programming tool. Hence, R helps in data importing and cleaning, depending on what manner of strategy you are using on.

R is best for data Science because it gives a broad variety of statistics. In addition, R provides the environment for statistical computing and design.

Better visualization since it has packages like plot3D, boxplots, barplots, shiny dashboard.

Feature Selection

Better decision making

Very powerful for finance

Can handle large dataset

Consists of multithreaded math libraries

Simple to learn and use as there are predefined library that have to be installed and use directly.

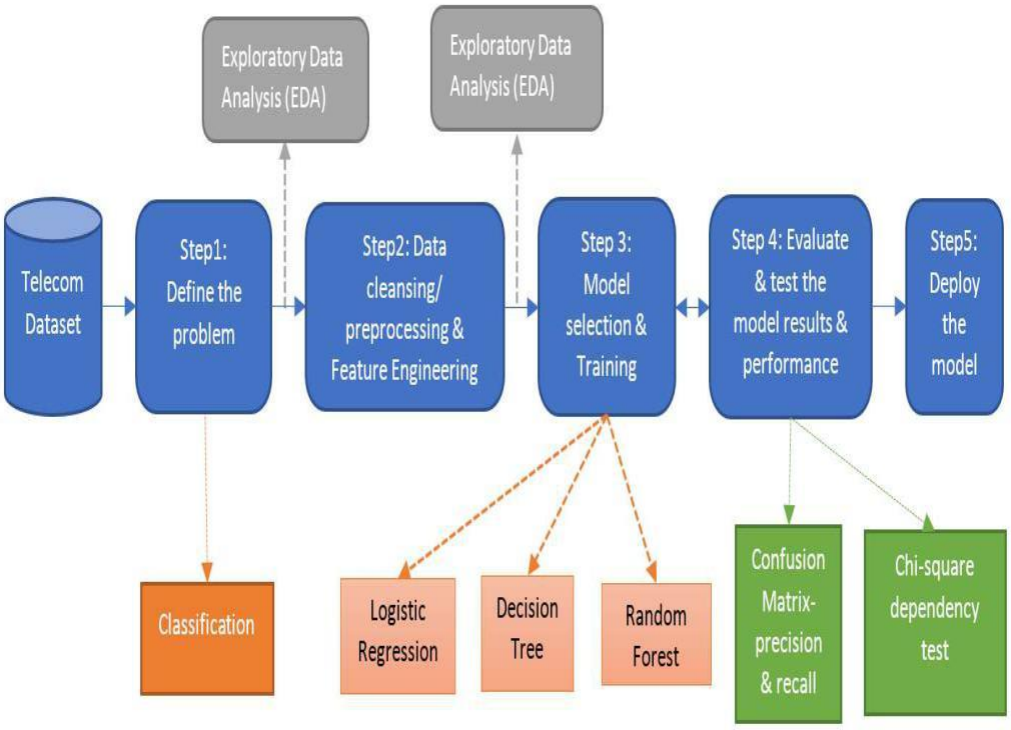


Fig 2: Customer Churn Model in R

1.4.3 Applications of R Programming

Many data analysts and research programmers use R because R is the most prevalent-language.

Hence, we use R as a fundamental tool for finance.

Many quantitative analysts use R as their programming tool. Hence, R helps in data importing and cleaning, depending on what manner of strategy you are using on.

R is best for data Science because it gives a broad variety of statistics.

In addition, R provides the environment for statistical computing and design. Rather R considers as an alternate execution of S.

1.4.4 Job Roles in R

Jobs in R Careers are not only being offered by IT companies but all types of companies are hiring High paid R candidates including-

Financial firms

Retail Organisation

Banks

Healthcare organizations etc.

There is a huge demand for R programmers among start-ups. Companies have several R job openings with various positions like:

R data scientist

Data-Scientist (IT)

Analyst manager

Senior data analyst

Business analyst

Analyst consultant

1.4.5 List of Companies Using R

So, below is the list of companies using R for analytics –

- **TechCrunch**
- **Google**
- **Facebook**
- **Genpact**
- **Bing**

- **ANZ**
- **The New York Times**
- **Thomas Cook**
- **Accenture**
- **Mozilla**
- **Novartis**
- **Merck**

1.4.6 Use Cases of R Language

- **Facebook**

Basically, Facebook uses R to update status and its social network graph. Also, for predicting colleague interactions with R, Facebook uses it.

- **Ford Motor Company**

As Ford relies on Hadoop. Also, R for statistical analysis and data-driven decision support.

- **Google**

Basically, Google uses R to calculate ROI on advertising campaigns, to predict economic activity. Also, to improve the efficiency of online advertising.

- **Microsoft**

Microsoft uses R for the Xbox matchmaking service. Also, as a statistical engine within the Azure ML framework.

- **Mozilla**

Generally, it is the foundation behind the Firefox web browser, uses R to visualize Web activity.

- **Twitter**

R is part of Twitter's Data-Science toolbox for sophisticated statistical modelling.

1.5 HARDWARE AND SOFTWARE SPECIFICATIONS

According to R specs page, the minimum requirements for running R Studio are:

1.5.1 Hardware

Table 1: Requirements for Hardware

Component	Specification
Processor	Core i5-5th Generation
RAM	4GB-16GB
HDD	500GB
Operating System	Windows 7/8/10

1.5.2 Software

Table 2: Requirements for Software

Programming Language	R version 3.5.3
Ecosystem	R Platform, MS Excel,
IDE	RStudio version 3.5.3
Text Editor	Notepad++ version 7.6.2

2. LITERATURE SURVEY

2.1 EXISTING WORK

Many Data Analytic and Visualization tools have been implemented during the past decade. Each tool has its own advantages and disadvantages. Some of these are R, Big-Data and Hadoop etc. Data analytics is the science of analyzing raw data in order to make conclusions about that information. Many of the techniques and processes of data analytics have been automated into mechanical processes and algorithms that work over raw data for human consumption. Data Visualization is an art of presenting data using perfect blend of colors, dimensions and labels. It is an inevitable aspect of Business Analytics. As more and more sources of data are getting discovered, business managers embrace data analytics and visualization software's to analyze trends and make quick decisions. We live in a time where data is all around us. Being a data-driven organization starts with understanding your data. Whether you're talking about an IT specialist, a CIO, or a project manager, in today's digital age, all business users need to be able to access and understand data. This methodology is an approach to data that supports business success and ensures that everyone within an organization is empowered to make the most of the information in front of them by understanding data in a seamless, interactive way. To develop a real modern business environment within your organization, you must implement the discovery of data so that you can remain relevant, successful, and facilitate a data-driven culture.

R programming language is best for statistical, data analysis and machine learning. By using this language, we can create objects, functions, and packages. We can use it anywhere. It's platform- independent, so we can apply it to all operating system. It's free, so anyone can install it in any organization without purchasing a license.

R is open source. Thus, Google is utilizing R programming as it is a suitable language.

By using R, we can create any form of statistics and data manipulation.

R, SAS, and SPSS are three statistical languages. Of these three statistical languages, R is the only an open source. SAS is the most important private software business in the world. SPSS is now overseen by IBM. R Programming is extensible and hence, R

groups are noted for its energetic contributions. Lots of Rs typical features can be written in R itself and hence, R has gotten faster over time and serves as a glue language.

2.2 PROPOSED METHOD

To identify the customers, we need to have a database with data about the previous customers that churned. Using this data, we develop a model which identifies customers that have a profile lose to the ones that already left. To simulate an experiment where we want to predict if our customers will churn, we need to work with a partitioned database. The database has 2 parts, one part will be the training set. This will be used to create the model. The second part will be the testing set which will be used to evaluate our model. In this case we know customer answers from the testing dataset so we can compare the model prediction with the true answers. Nevertheless, in reality, we don't know what will be the true answers. So, we have to target mainly customers with high probability to churn. This probability is given by our model.

Fig 3: Proposed Methodology in Churn Prediction

We can estimate our train with 3 different models.

1)Decision Tree

2)Random Forest

3)Logistic Regression Model

2.2.1 Confusion Matrix:

A confusion matrix is a table that is often used to describe the performance of a classification model (or “classifier”) on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm.

2.2.2 Accuracy:

Classification Rate/Accuracy: Classification Rate or Accuracy is given by the relation:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

True Positive (TP): Observation is positive, and is predicted to be positive**.**
False Negative (FN): Observation is positive, but is predicted negative.
True Negative (TN): Observation is negative, and is predicted to be negative.
False Positive (FP): Observation is negative, but is predicted positive.

2.2.3 Kappa Statistics:

Cohen's kappa. Cohen's kappa coefficient (κ) is a statistic which measures inter-rater agreement for qualitative (categorical) items. It is generally thought to be a more robust measure than simple percent agreement calculation, as κ takes into account the possibility of the agreement occurring by chance.

Value of Kappa	Level of Agreement	% of Data that are Reliable
0-.20	None	0-4%
.21-.39	Minimal	4-15%
.40-.59	Weak	15-35%
.60-.79	Moderate	35-63%
.80-.90	Strong	64-81%
Above .90	Almost Perfect	82-100%

Fig 4: kappa statistics

2.2.4 Decision Tree:

Decision tree is a graph to represent choices and their results in form of a tree. The nodes in the graph represent an event or choice and the edges of the graph represent the decision rules or conditions.

Syntax: ctree(formula, data)

formula is a formula describing the predictor and response variables.

data is the name of the data set used.

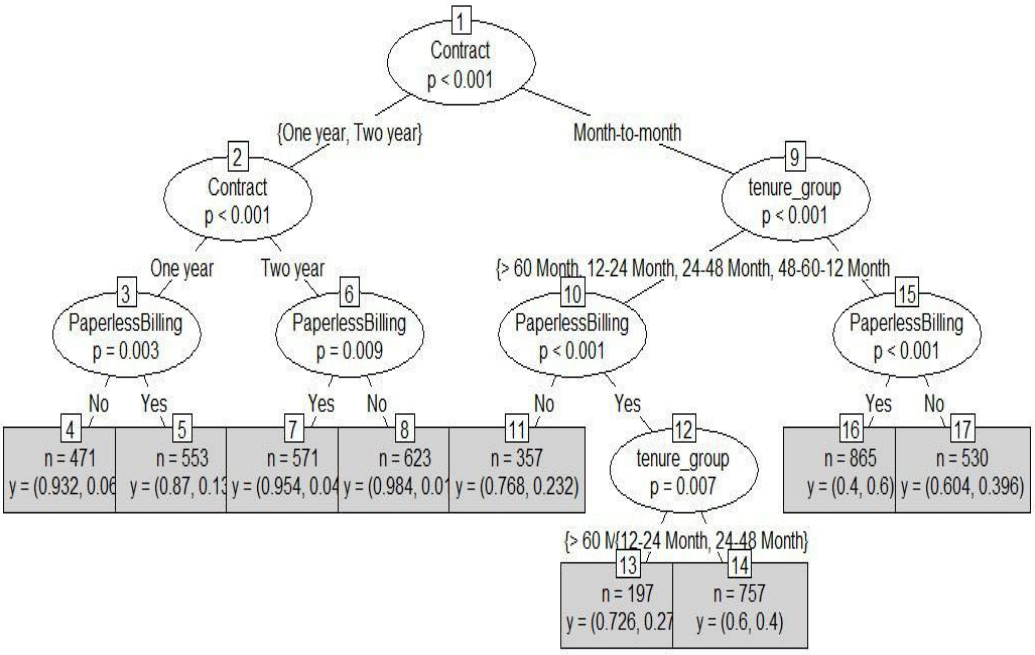


Fig 5: Decision Tree Model Analysis

In the graphic above, we see the decision tree that groups observations by their variable values. A decision tree has 2 mains components, leaves and nodes.

Leaves represent a group of observation. For each leaf, an answer is given, “Yes” or “No”. Below these answers, figures represent the percentage of churn in a leaf and finally we see the percentage of total observations in the leaf.

Nodes show which variable where used to separate a leaf in two sub-leaves

Contract is the main variable in the churn decision. It makes sense because it is harder to change telecom providers if customers have a long-term contract than a month-to-month contract.

2.2.5 Logistic Regression Model:

The Logistic Regression model is a model in which dependent variable has categorical values such as True/False or 0/1. It measures the probability of a binary response as the value of response variable based on the mathematical equation relating it with the predictor variables. The **glm()** function is to create the regression model.

Syntax: glm(formula, data, family)

2.2.6 Random Forest:

In the random forest approach, a large number of decision trees are created. Every observation is given into every decision tree. An error estimate is made for the cases which were not used while building the tree. The R package "**randomForest**" is used to create random forests.

Syntax: randomForest(formula, data)

Formula: is a formula describing the predictor and response variables.

Data: is the name of the data set used.

2.3 DATA MINING TECHNIQUES

The process of reducing, analyzing the patterns, predicting the hidden and useful required information from large Database is known as Data Mining. Association rule mining, clustering, classification and regression forms the four techniques used by data mining.

In Data mining new rules and patterns can be discovered by the system known as discovery oriented and system can also check the user ‘s hypothesis called verification oriented. It helps in taking knowledge-driven decisions and for predicting the future trends of the business.

2.3.1 Tool Used: A Revolution Analytics Tool - R

In the past few years, the fast-emerging requirements from both academia and industry has helped R programming language to emerge as one of the necessary tools for visualization, computational statistics and data science. R is most popular in field of data science and important in Finance and analytics- driven companies.

R virtually consists all the possible statistical models, data manipulation and charts that could ever be required by a modern-day scientist. One can easily use the best reviewed methods from leading researchers in field of Data Science without any cost. It provides a large collection of graphical and statistical techniques, consisting of modelling (linear and non-linear), statistical tests, time-series, classification, clustering, etc.

R helps in representing complex data as beautiful and unique data visualizations. Evaluation of result in R is very much easier as we do not have to remember any clicks or steps, it is simply a programming language designed specifically for data analysis that also has the capability to use mix and match models for best results.

As R is supported by a large community worldwide, solution to the errors and code is available freely. Its source code is written in C, Fortran and R. R is easily extensible through functions and extensions, and the R community is noted for its active contributions in terms of packages. R is an open source and can be extended easily as individuals using it can contribute in its growth. Dynamic and static graphics are available through additional packages. R can easily deal with complex and large datasets.

The libraries and packages of R that are being used in this paper are: RWeka, ggplot2, rpart, rJava, class.

3. SYSTEM DESIGN

3.1 DESIGN STEPS

The study of predicting which persons are going to churn in advance will help the telecommunication industry and the CRM department to identify which persons are going to leave the network. The problem of our work discussed is the classification problem i.e. to classify each subscriber as potential churner or potential non churner. The framework discussed below is based on the Knowledge Discovery Data (KDD) process. Our framework consists of the following five modules:

1. **Data Acquisition**
2. **Data Preparation**
3. **Data Pre-processing**
4. **Data Extraction**
5. **Decision Making**

Data Acquisition: Acquiring data from the tele set industry is a big task because of the fear of misusing it. The data set for this study acquired from the KDD from Kaggle. It is used to analyse the marketing tendency of customers from the large databases from the French Telecom company.

Data Preparation: Since the dataset acquired cannot be applied directly to the churn prediction models, so aggregation of data is required where new variables are added to the existing variables by viewing the periodic usage behaviour of the customers. These variables are very important in predicting the behaviour of customers in advance as they contain critical information used by the prediction models.

Data Pre-processing: Data pre-processing is the most important phase in prediction models as the data consists of ambiguities, errors, redundancy which needs to be cleaned beforehand. The data gathered from multiple sources first is aggregated and then cleaned as the complete data collected is not suitable for modelling purposes. The

records with unique values do not have any significance as they do not contribute much in predictive modelling. Fields with too many null values also need to be discarded.

Data Extraction: The attributes are identified for classifying process. In our work, we have worked with numerical and categorical values.

Decision: The rule set will let the subscribers identify and classify in the different categories of churners and non-churners by setting a particular threshold value.

The framework design used in our work is given in Figure 3. We have taken dataset which consists of total 3333 attributes.

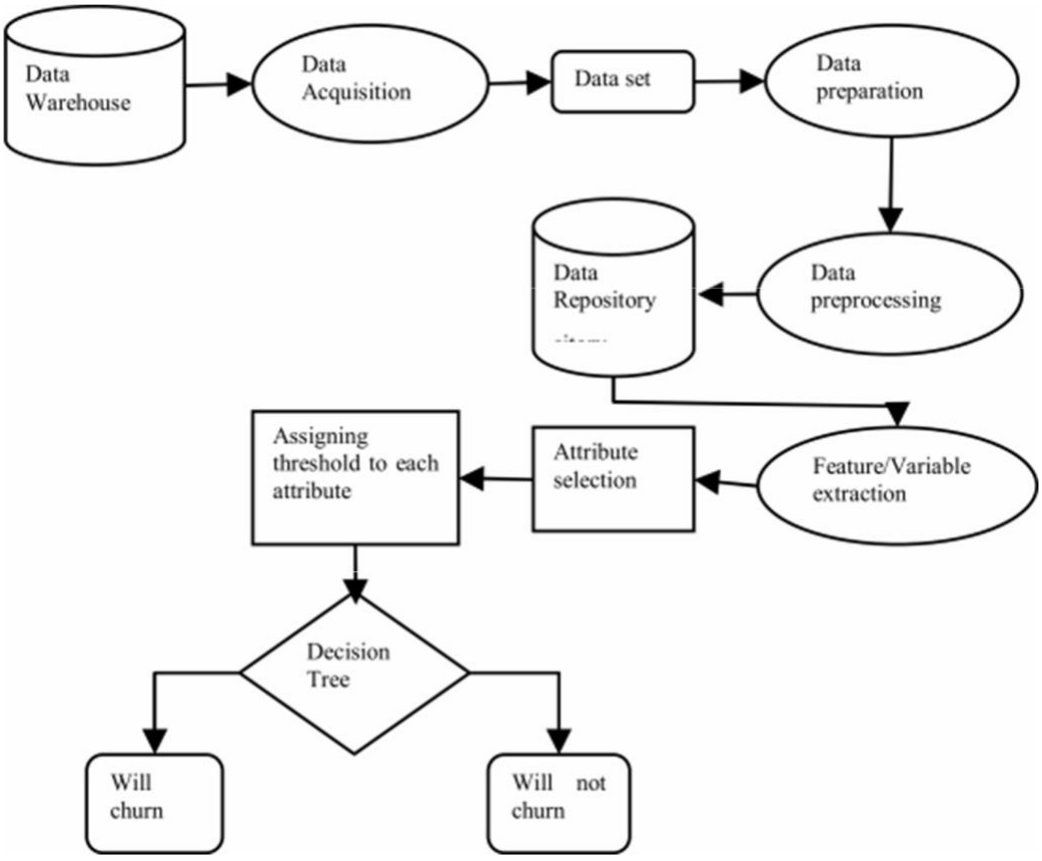


Fig 6: Churn Prediction Framework

3.2 MANAGING CHURNS

Churn management is very important for reducing churns as acquiring a new customer is more expensive than retaining the existing ones. Churn rate is the measurement for the number of customers moving out and in during a specific period of time. If the reason for churning is known, the providers can then improve their services to fulfil the needs of the customers.

Churns can be reduced by analyzing the past history of the potential customers systematically. Large amount of information is maintained by telecom companies for each of their customers that keeps on changing rapidly due to competitive environment. This information includes the details about billing, calls and network data. The huge availability of information arises the scope of using Data mining techniques in the telecom database. The information available can be analyzed in different perspectives to provide various ways to the operators to predict and reduce churning. Only the relevant details are used in analysis which contribute to the study from the information given.

Data mining techniques are used for discovering the interesting patterns within data. One of the most common data mining technique is Classification, its aim is to classify unknown cases based on the set of known examples into one of the possible classes. Here, in case of telecom churn, Classification helps learn to predict whether a customer will churn or not based on customer's data stored in database.

3.3 PROPOSED ANALYSIS FOR BUSINESS ORGANIZATIONS

To identify the customers, we need to have a database with data about the previous customers that churned. Using this data, we develop a model which identifies customers that have a profile lose to the ones that already left. To simulate an experiment where we want to predict if our customers will churn, we need to work with a partitioned database. The database has 2 parts, one part will be the training set. This will be used to create the model. The second part will be the testing set which will be used to evaluate our model. We have to target mainly customers with high probability to churn.

3.4 Algorithm Incorporated with R and Pseudo Code

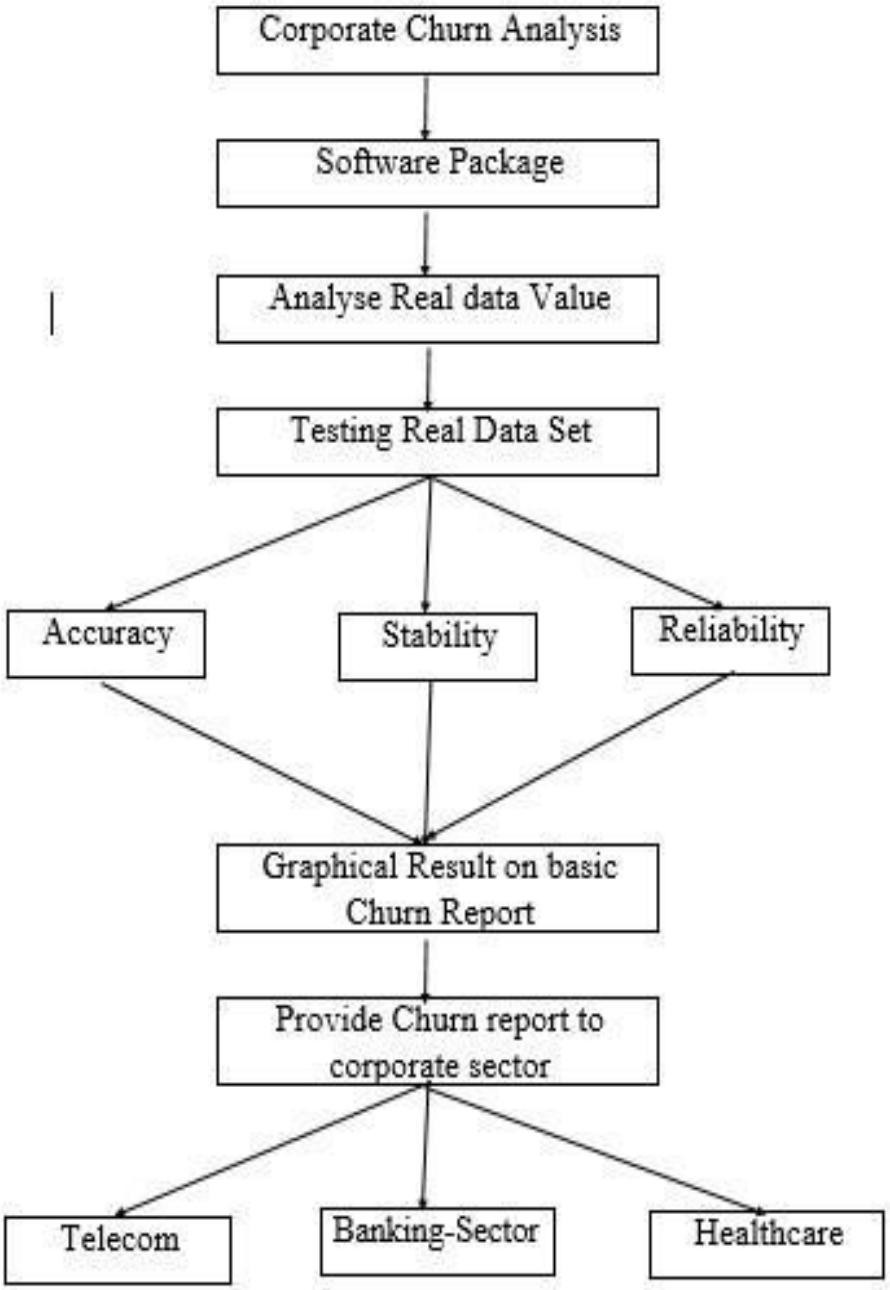


Fig 7: Analysis and Testing Data on Customer Churn

3.5 Pseudo Code of Statistical Churn Analysis

```
train = read.csv('Telecom_Train.csv')
train_df = train[,-1]
head(train_df)

test = read.csv('Telecom_Test.csv')
test_df = test[,-1]
head(test_df)

str(train_df)
summary(train_df)


traindf = train_df[,c(2,6:19)]
testdf = test_df[,c(2,6:19)]
par(mfrow=c(1,2))
for(i in 1:ncol(traindf)){
  boxplot(traindf[,i],col="green",border="brown",notch=TRUE, main = names(traindf)[i])
}
train_cat = train_df[, c(-2,-6:-19)]
test_cat = test_df[, c(-2,-6:-19)]
head(train_cat)
head(test_cat)
par(mfrow = c(2,1))


barplot(table(train_cat$state),main="State",col="blue",border="black")
barplot(table(train_cat$area_code),main="Area_Code",col="green",border="black")



#*********************************************
# Decision Tree model- method C5.0Cost
fit_c50Cost<- train(churn~.,data=train_df[,-1],
                    trControl=train_control,
                    method='C5.0Cost')

fit_c50Cost
predictions <- predict(fit_c50Cost,test_df)
pred = cbind(test_df,predictions)
confusionMatrix(pred$churn,pred$predictions)
varImp(fit_c50Cost)
# Accuracy =0.9574  and Kappa =0.7991
```

3.6 TESTING PROCESS

The aim of this study is to identify the different algorithm used for churn prediction. Once churns are identified company has to take further action to prevent the churn. The managerial department of that company has to use such identified churns and put efforts as how much revenue a service provider is going to get over are to be made to retain period of customer remain. In general, the customer lifetime value is highly connected with the customer decision to stay back. The customer lifetime value can be combined with the churn prediction to reduce the cost for making a excessive retention effort (false positives) and the cost of losing a customer because the model not accurately predict churn (false negatives).

Table 3: Churn Prediction Categories

	Actual Churners	**Actual Non-Churners**
Prediction Churners	**True Positive**	**False Positive**
Prediction Non-Churners	**True Negative**	**False Negative**

3.6.1 Testing and validation of the model on real customers

When the model is ready and it is ensured that it takes into account all the special needs of a particular business and its customers, the testing period starts. It normally takes up to a few months. The model is fine-tuned according to the results. Such custom-built models have a solid advantage compared to automatically generated models — they stay very flexible and can be developed according to each company's growing demands.

3.6.2 Data Set Used

Table 4: Dataset Attributes

S.No.	Attribute name
1	**State**
2	**Account. Length**
3	**Area. Code**
4	**Phone**
5	**Int .l .Plan**
6	**VMail.Plan**
7	**VMail.Message**
8	**Day.Mins**
9	**Day.Calls**
10	**Day.Charge**
11	**Eve.Mins**
12	**Eve.Calls**
13	**Eve.Charge**
14	**Night.Mins**
15	**Night.Calls**
16	**Night.Charge**
17	**Intl.Mins**
18	**Intl.Calls**

19	**Intl.Charge**
20	**CustServ.Calls**
21	**Churn.**

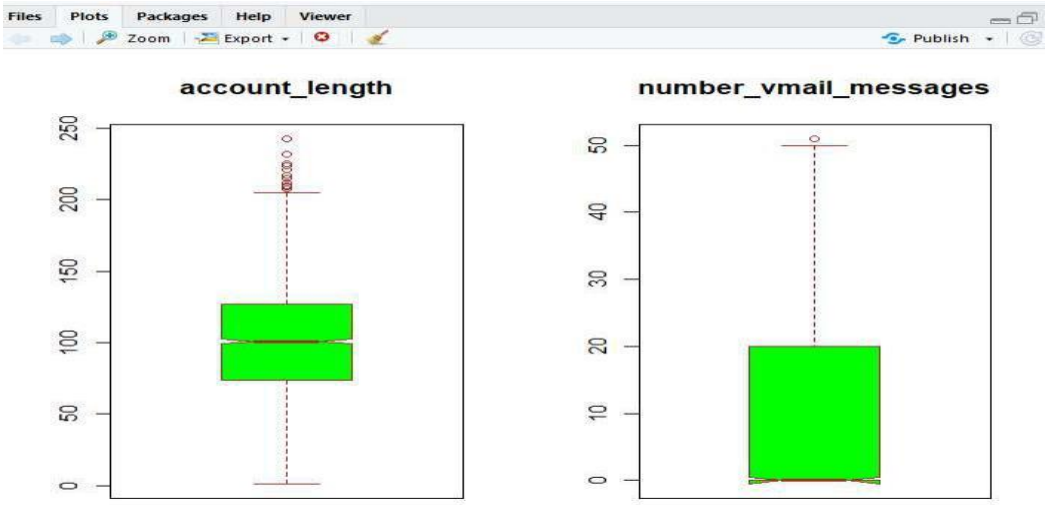
3.6.3 Description of complete data Set

```
> str(churn)
'data.frame':   3333 obs. of  21 variables:
 $ State         : Factor w/ 51 levels "AK","AL","AR",..: 17 36 32 36 37 2 20 2$
 $ Account.Length: int  128 107 137 84 75 118 121 147 117 141 ...
 $ Area.Code     : int  415 415 415 408 415 510 510 415 408 415 ...
 $ Phone         : Factor w/ 3333 levels "327-1058","327-1319",..: 1927 1576 11$
 $ Int.l.Plan    : Factor w/ 2 levels "no","yes": 1 1 1 2 2 2 1 2 1 2 ...
 $ VMail.Plan    : Factor w/ 2 levels "no","yes": 2 2 1 1 1 1 2 1 1 2 ...
 $ VMail.Message : int  25 26 0 0 0 0 24 0 0 37 ...
 $ Day.Mins      : num  265 162 243 299 167 ...
 $ Day.Calls     : int  110 123 114 71 113 98 88 79 97 84 ...
 $ Day.Charge    : num  45.1 27.5 41.4 50.9 28.3 ...
 $ Eve.Mins      : num  197.4 195.5 121.2 61.9 148.3 ...
 $ Eve.Calls     : int  99 103 110 88 122 101 108 94 80 111 ...
 $ Eve.Charge    : num  16.78 16.62 10.3 5.26 12.61 ...
 $ Night.Mins    : num  245 254 163 197 187 ...
 $ Night.Calls   : int  91 103 104 89 121 118 118 96 90 97 ...
 $ Night.Charge  : num  11.01 11.45 7.32 8.86 8.41 ...
 $ Intl.Mins     : num  10 13.7 12.2 6.6 10.1 6.3 7.5 7.1 8.7 11.2 ...
 $ Intl.Calls    : int  3 3 5 7 3 6 7 6 4 5 ...
 $ Intl.Charge   : num  2.7 3.7 3.29 1.78 2.73 1.7 2.03 1.92 2.35 3.02 ...
 $ CustServ.Calls: int  1 1 0 2 3 0 3 0 1 0 ...
 $ Churn.        : Factor w/ 2 levels "False.","True.": 1 1 1 1 1 1 1 1 1 1 ...
```

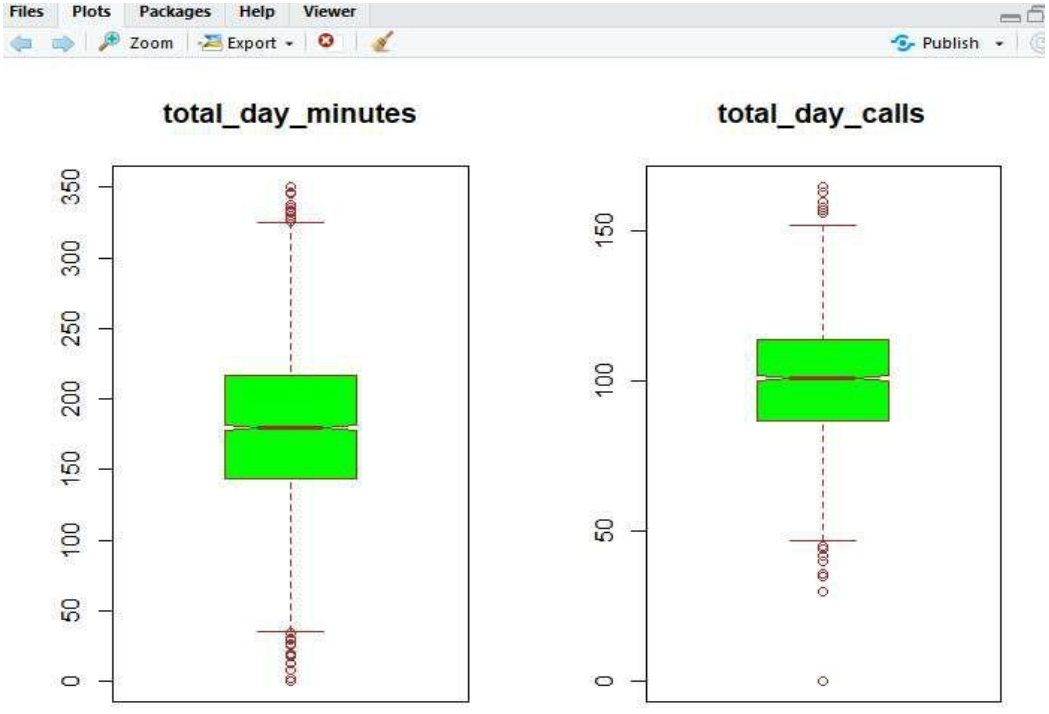
4. RESULTS/SNAPSHOTS OF THE PROJECT

4.1 ANALYSIS OF NON-CATEGORICAL DATASET

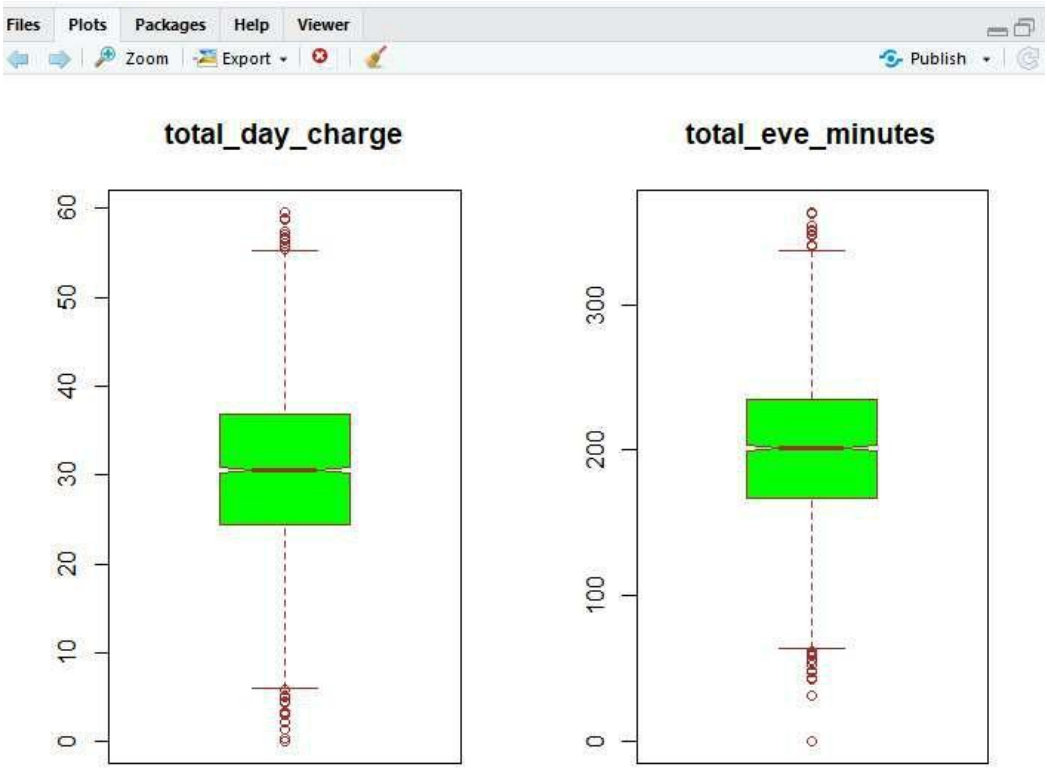
4.1.1



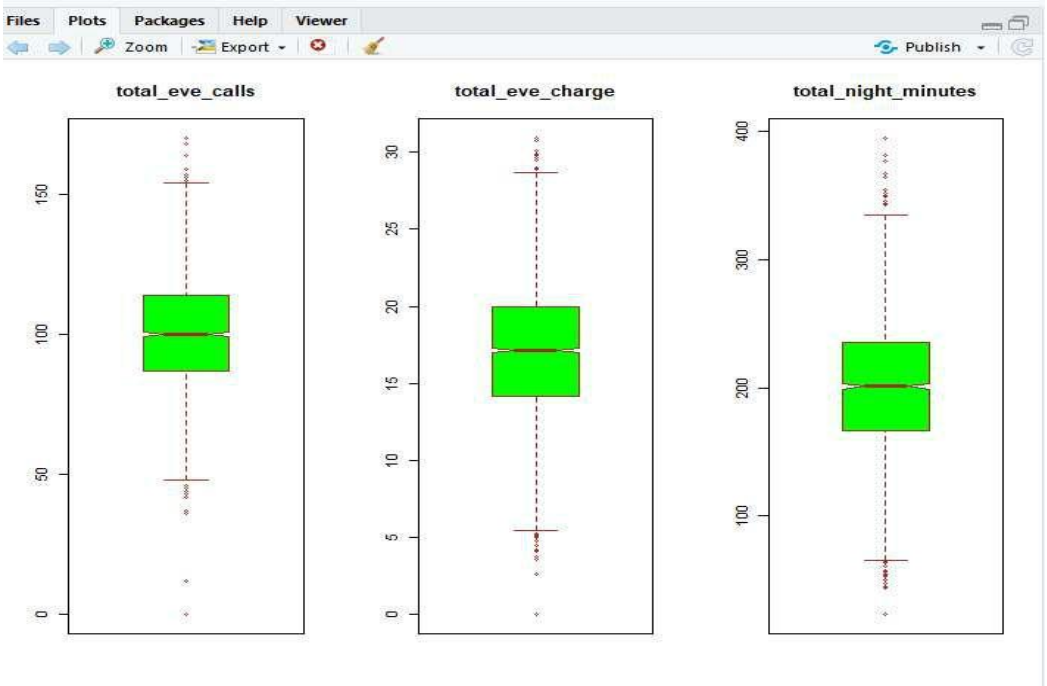
4.1.2



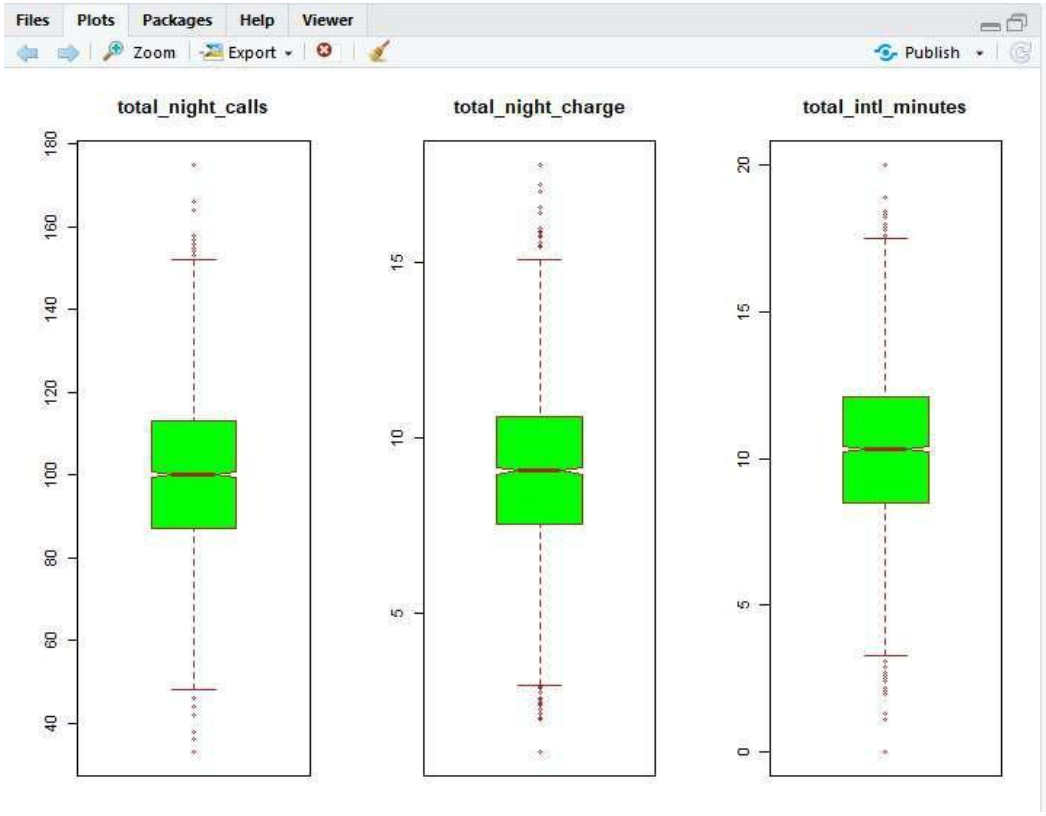
4.1.3


Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
total_day_charge
total_eve_minutes
0
10
20
30
40
50
60
0
100
200
300

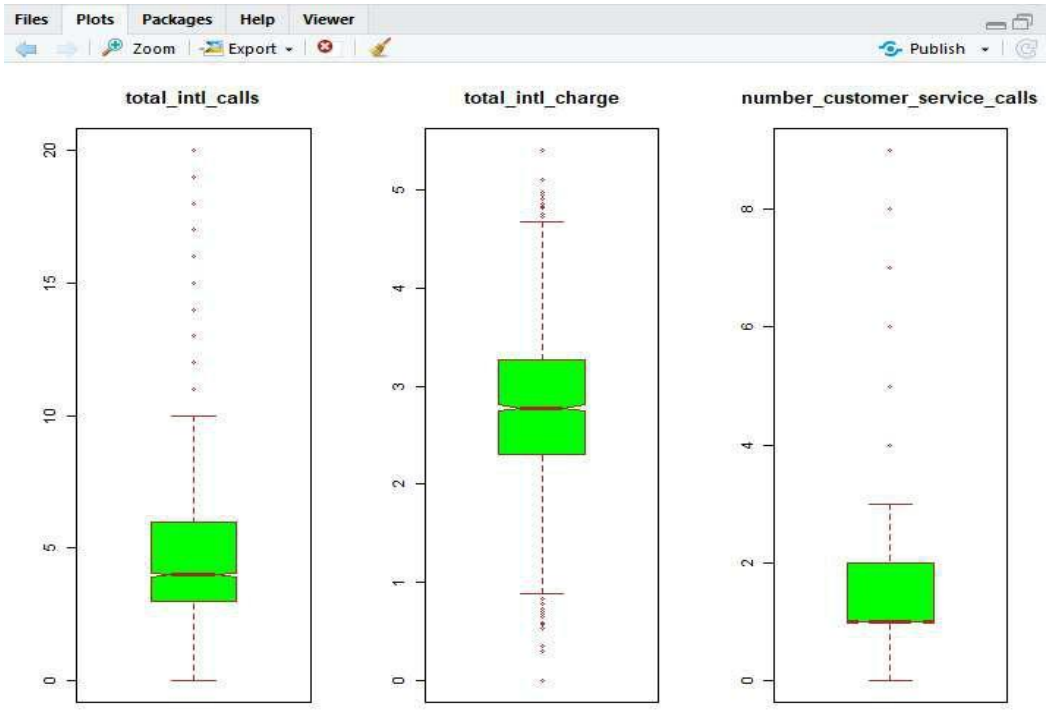
4.1.4


Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
total_eve_calls
total_eve_charge
total_night_minutes
0
50
100
150
0
5
10
15
20
25
30
100
200
300
400

4.1.5

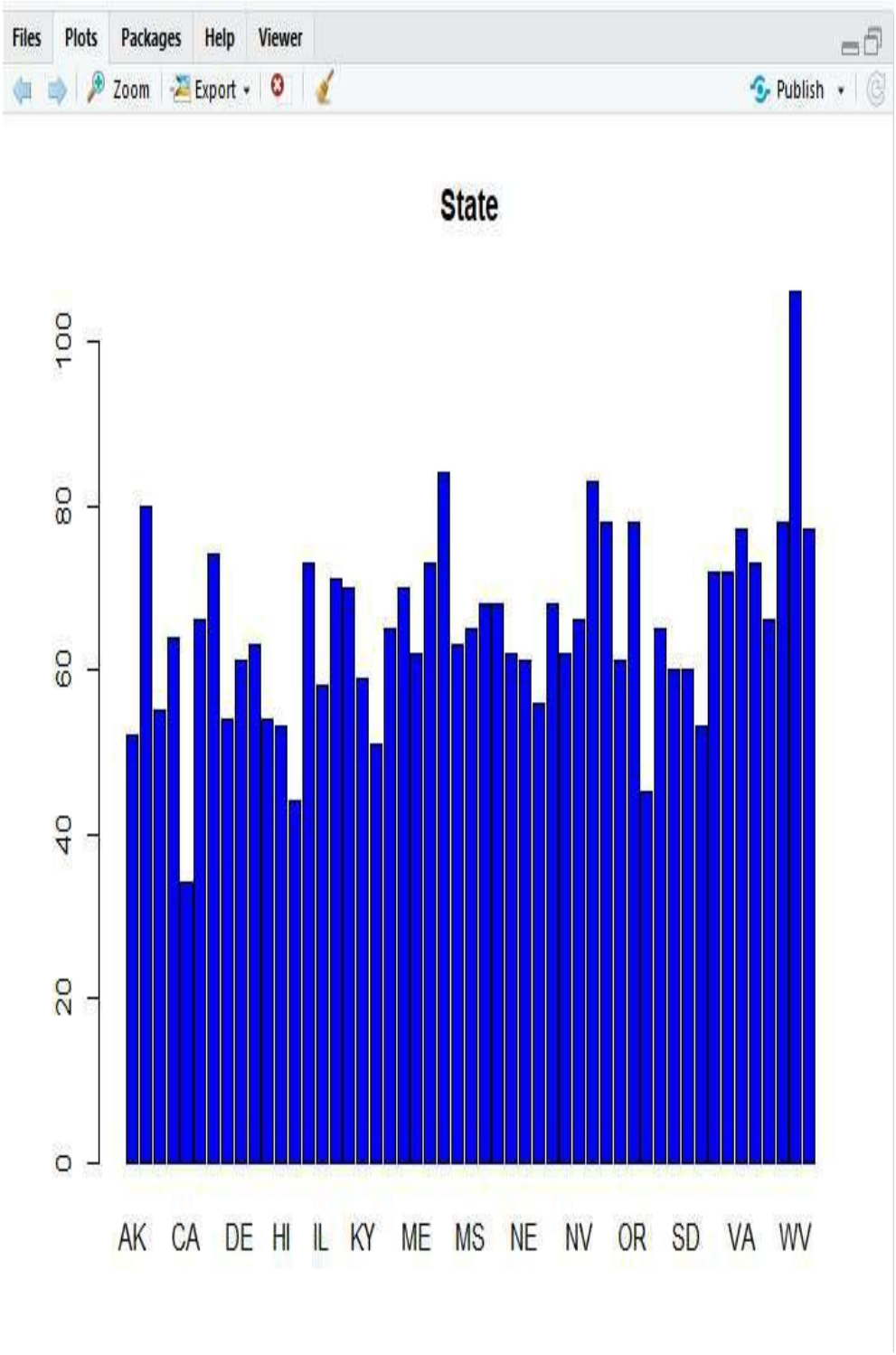

Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
total_night_calls
total_night_charge
total_intl_minutes
180
160
140
120
100
80
60
40
20
15
10
5
0

4.1.6

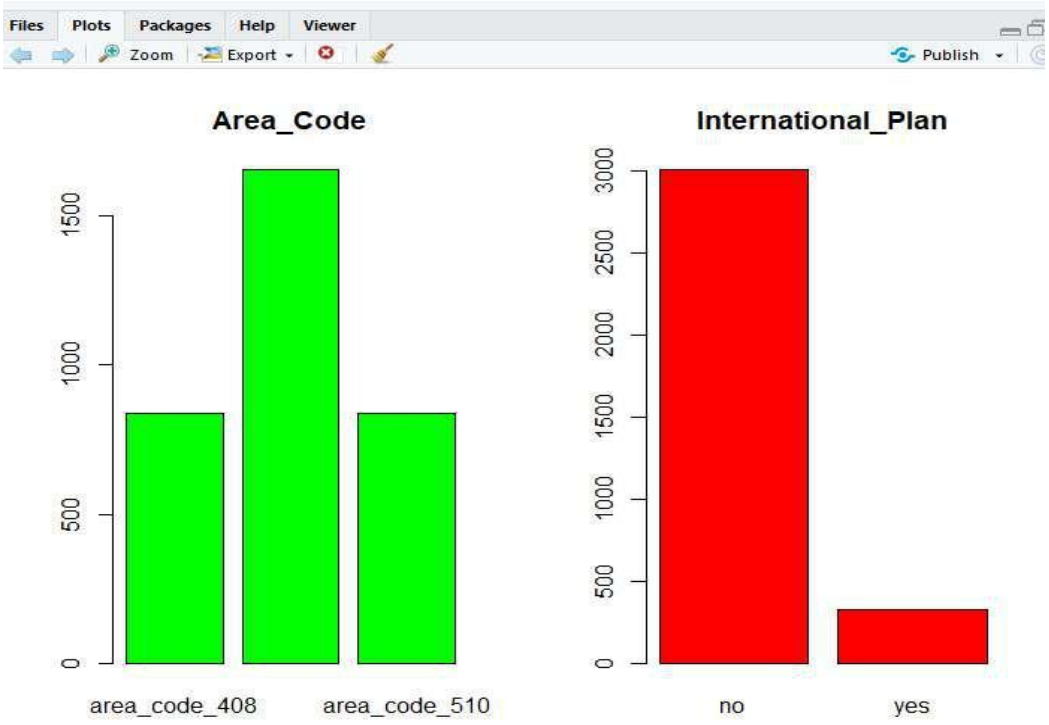

Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
total_intl_calls
total_intl_charge
number_customer_service_calls
20
15
10
5
0
5
4
3
2
1
0
8
6
4
2
0

4.2 ANALYSIS OF CATEGORICAL DATASET

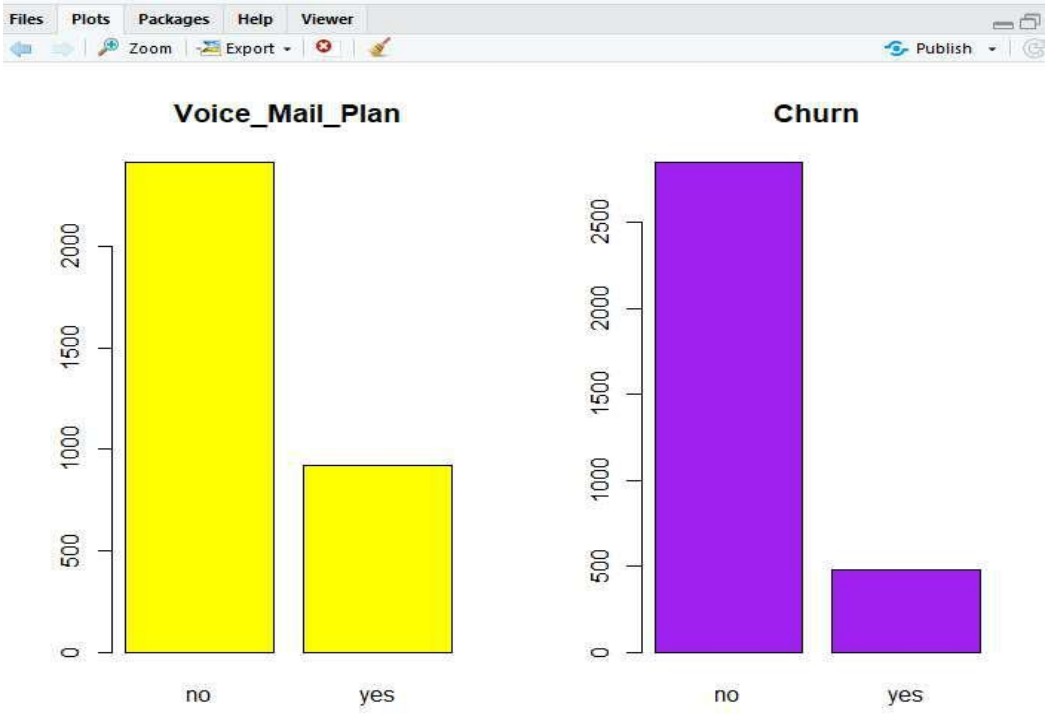
4.2.1



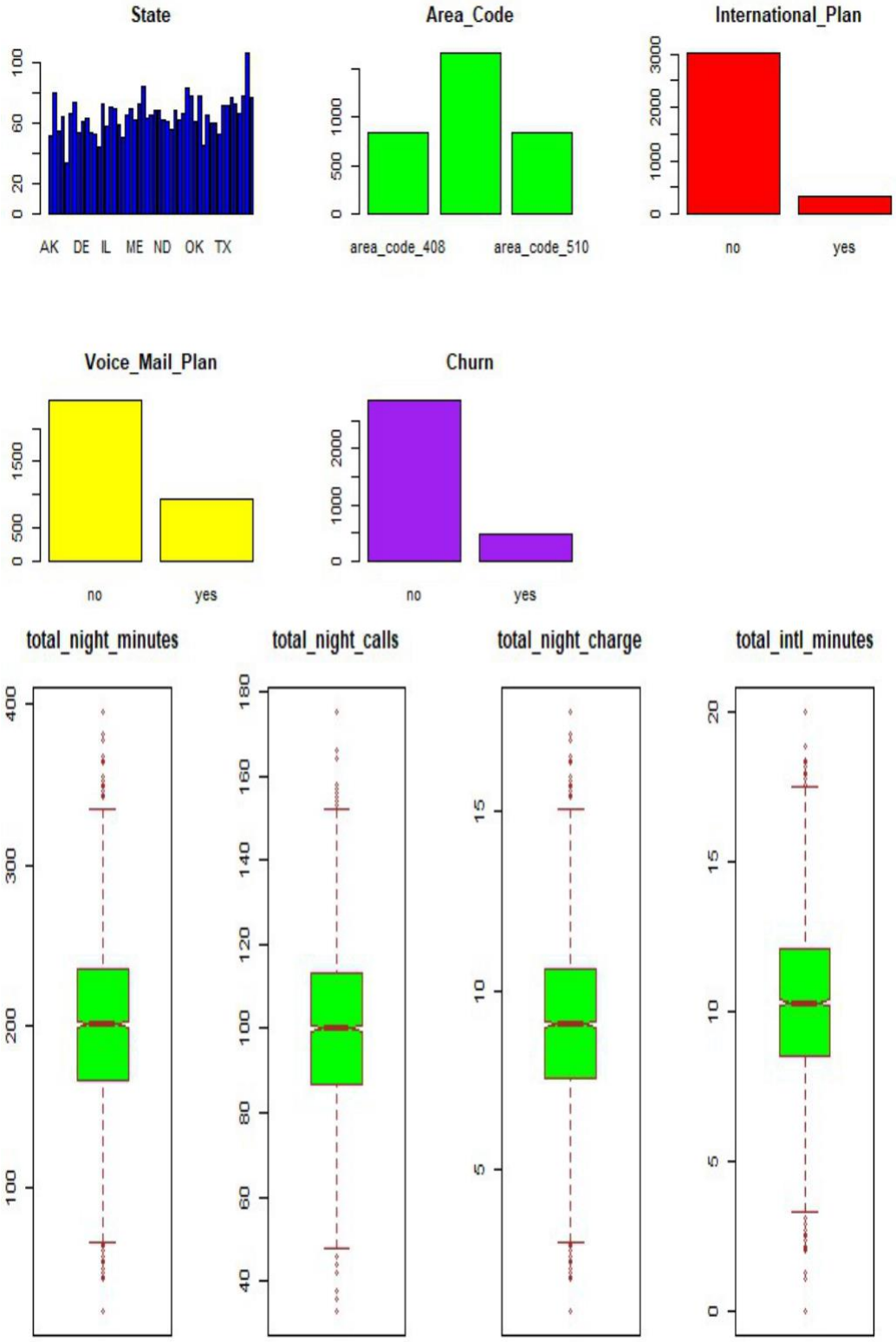
4.2.2


Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
Area_Code
0
500
1000
1500
area_code_408
area_code_510
International_Plan
0
500
1000
1500
2000
2500
3000
no
yes

4.2.3


Files
Plots
Packages
Help
Viewer
Zoom
Export
Publish
Voice_Mail_Plan
0
500
1000
1500
2000
no
yes
Churn
0
500
1000
1500
2000
2500
no
yes

4.3 Overall Overview of Categorical and Non-Categorical Data



4.4 COMPARISON OF VARIOUS PREDICTION MODEL USED IN CHURN ANALYSIS

Table 5: Various Prediction Model Comparison

Model Used	Metric	value
Logistic	#Accuracy	0.8368
	#Kappa	0.3065
Rpart	#Accuracy	0.8872
	#Kappa	0.3413
C50	#Accuracy	0.9568
	#Kappa	0.7942
GLM	#Accuracy	0.8716
	#Kappa	0.2305
bstTree	#Accuracy	0.9442
	#Kappa	0.7182
c5.0 cost	#Accuracy	0.9574
	#Kappa	0.7991
c5.0 rules	#Accuracy	0.9454
	#Kappa	0.7371
treebag	#Accuracy	0.952
	#Kappa	0.7732
xgbTree	#Accuracy	0.9502
	#Kappa	0.7592

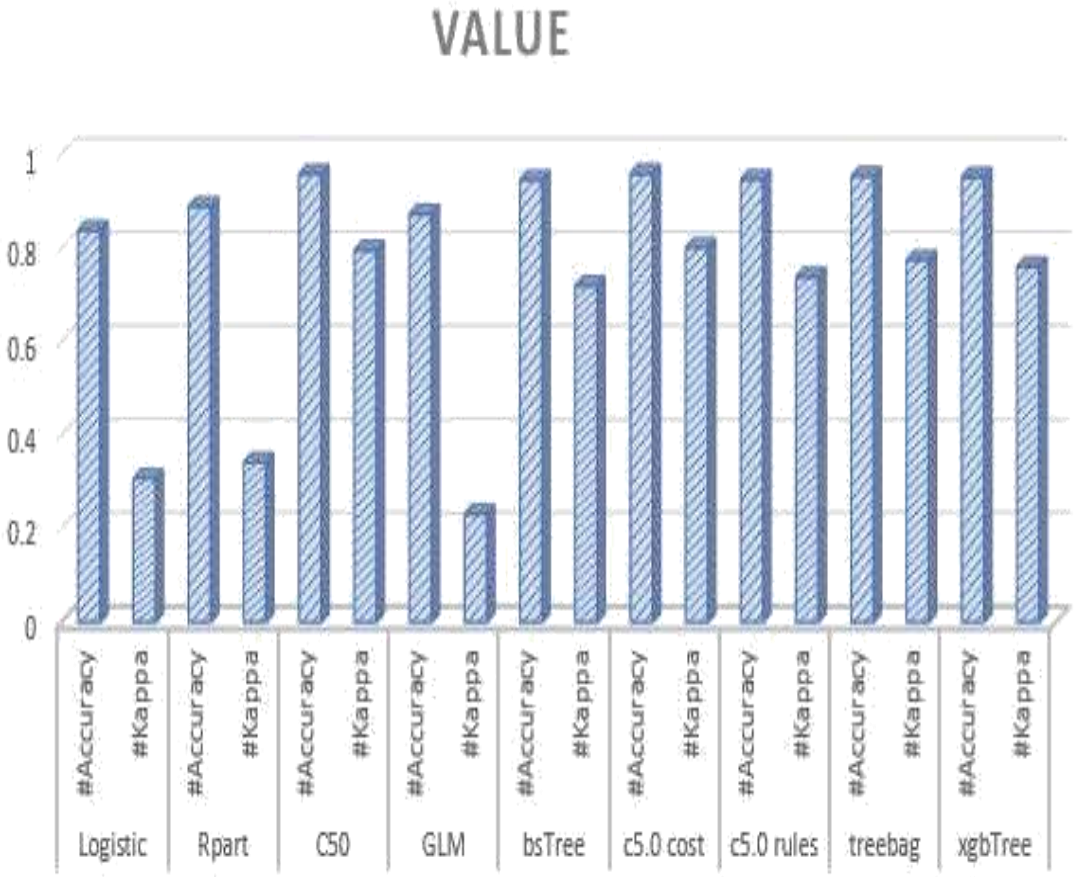


Fig 11: Graphical comparison of various Prediction Models

5. CONCLUSION

Telecommunication industry has suffered from high churn rates and immense churning loss. Although the business loss is unavoidable, but still churn can be managed and kept in an acceptable level. Good methods need to be developed and existing methods have to be enhanced to prevent the telecommunication industry to face challenges. In this project we discussed the various prediction models and also compared the quality measures of prediction models like regression analysis, decision trees. The churn prediction model helps in better decision making. After analyzing various predictive models, we found that the accuracy achieved with decision tree using C5.0 Cost method is far much higher (above 86 %) than the logistic regression technique (less than 80%) which clearly states that decision tree is an efficient technique.

6. FUTURE SCOPE OF THE PROJECT

The future scope of this project will use hybrid classification techniques to point out existing association between churn prediction and customer lifetime value. The retention policies need to be considered by selecting appropriate variables from the dataset. The passive and the dynamic nature of the industry ensure that data mining has become increasingly significant aspect in the telecommunication industry prospect. So, this project will let the telecommunications company to measure the churn of their customers and help them to make better relationships with their existing customers.

The future scope of this project will be to acquire 100% accuracy with good kappa value so that this may also be used in medical fields. The proposed model can be further enhanced, if the processes are working in parallel.

7. REFERENCES

1. Tutorials at https://www.w3schools.in/r/R
2. R User Guide at http://r-tutorials.com/
3. Online Resources https://www.guru99.com/r-tutorial.html
4. Online Resource for datasets at

   https://www.kaggle.com/jherfordsasm/telecom-customer-churn-datasets-traintest
5. R Programming for Data Science by Roger Peng
6. R for Everyone: Advanced Analytics and Graphics (2nd Edition) (Addison-Wesley Data & Analytics Series).